Precise Generation of Conformational Ensembles for Intrinsically Disordered Proteins Using Fine-tuned Diffusion Models

Junjie Zhu[1,#], Zhengxin Li[1,#], Bo Zhang[1], Zhuoqi Zheng[1], Bozitao Zhong[1], Jie Bai[1], Taifeng Wang[2,*], Ting Wei[1,*], Jianyi Yang[3,*], Hai-Feng Chen[1,*]

[1]State Key Laboratory of Microbial Metabolism, Department of Bioinformatics and Biostatistics, SJTU-Yale Joint Center for Biostatistics, National Experimental Teaching Center for Life Sciences and Biotechnology, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, 200240, Shanghai, China.

[2]BioMap,10 Beilun Industrial Park, Yongteng North Road, 100080, Beijing, China.

[3]MOE Frontiers Science Center for Nonlinear Expectations, Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, 266237, Qingdao, China.

***Corresponding Author**
Email: haifengchen@sjtu.edu.cn; taifeng@biomap.com; yangjy@sdu.edu.cn

Notes

The authors declare that there is no conflict of interest.

Author Contributions

[#]These authors contributed equally to this work.

Abstract

Intrinsically disordered proteins (IDPs) play pivotal roles in various biological functions and are closely linked to many human diseases including cancer. Structural investigations of IDPs typically involve a combination of molecular dynamics (MD) simulations and experimental data to correct for intrinsic biases in simulation methods. However, these simulations are hindered by their high computational cost and a scarcity of experimental data, severely limiting their applicability. Despite the recent advancements in structure prediction for structured proteins, understanding the conformational properties of IDPs remains challenging partly due to the poor conservation of disordered protein sequences and limited experimental characterization. Here, we introduced IDPFold, a method capable of predicting IDP conformation ensembles directly from their sequences using fine-tuned diffusion models. IDPFold bypasses the need for Multiple Sequence Alignments (MSA) or experimental data, achieving accurate predictions of ensemble properties across numerous IDPs. By sampling conformations at the backbone level, IDPFold provides more detailed structural features and more precise property estimation compared to the state-of-the-art methods, and will help to reveal the disorder-function paradigm of IDPs.

Introduction

Intrinsically disordered proteins (IDPs) constitute a category of proteins with unstable structures under physiological conditions[1]. These proteins, accounting for over 40% of eukaryotic proteomes[1], are involved in various biological functions including signal transduction, molecular recognition, and cell cycle regulation[2–4]. IDPs are also closely associated with various significant diseases, such as cancer, Parkinson's disease, and acquired immunodeficiency syndrome (AIDS)[5–7]. Unlike structured proteins that possess one or a few stable conformations, IDPs exhibit transitions between multiple conformations with very low energy barriers, constantly fluctuating within a broad ensemble of structures under physiological conditions[8]. Consequently, deciphering the conformational ensemble of IDPs poses a significant challenge for experimental methods such as X-ray, cryo-electron microscopy and NMR[9–12].

Currently, molecular dynamics (MD) simulation is the most commonly used and effective tool for sampling conformational ensembles[13–15]. By iteratively sampling target molecular system based on first principles, an estimation of conformational ensemble is obtained from simulation. However, MD simulation incurs high sampling costs, and ensuring thorough sampling of the ensemble is challenging. Additionally, commonly used force fields for IDPs, such as ff03CMAP[16], ff14IDPs[17], ESFF1[18], still exhibit considerable errors in estimating local and global properties of IDPs during simulations[19,20]. Although we can adjust MD simulations to better align with the characteristics of IDPs using experimental data, only a few of these proteins have been experimentally characterized[21]. These limitations constrain the application of MD in predicting the conformational ensemble of IDPs.

On the other hand, numerous deep learning methods have been developed and widely applied in the field of structure prediction for structured protein, such as AlphaFold2 and ESMFold[22,23]. Simultaneously, in protein design tasks, various generative models, such as RFdiffusion and Chroma, have been utilized for

generating diverse protein backbones[24,25]. Therefore, it is natural to consider whether deep learning methods can be employed for rapid and accurate prediction of conformational ensembles of IDPs. However, the structures associated with IDPs are currently very sparse. The PDB database contains approximately 200,000 structural entries[26], while the Protein Ensemble Database only includes 553 protein ensemble data[27]. Moreover, IDPs often lack high-quality multiple sequence alignment (MSA) data. MSA has been proven effective only in functional studies concerning folded or bound states of IDPs, and provides little assistance in predicting more disordered conformations[28]. These challenges make it highly daunting to use deep learning for robustly predicting the conformational ensembles of IDPs.

We introduced IDPFold here for predicting IDP conformational ensembles directly from sequences based on a generative deep learning model. IDPFold utilized a protein language model to extract sequence information and further fed it into a structure generation module, enabling MSA-free conformation generation. To address the issue of insufficient data, we employed a hybrid dataset comprising crystal structures, NMR structures and MD trajectories for training and evaluation. IDPFold could generate IDP conformation ensembles at backbone level that is in better agreement with experimental observations than SOTA methods and samples conformational transitions from structured to disordered states, and provide insights for studying the correlation between structures and functions of IDPs.

Results

Overview of IDPFold

IDPFold employed a conditional diffusion model framework for generating protein conformation ensembles from sequences. This framework involved a forward diffusion process where noise is gradually added to real protein structures, and a reverse diffusion process where a deep learning network is used for denoising. By

integrating specific protein sequence features into the model during the reverse diffusion process, we generated conformation ensembles for specific proteins using this architecture. We constructed denoising networks using DenoisingIPA Blocks similar to the structural modules of AlphaFold2 to capture the chain-like structure within proteins and the rotational/translational state of each residue (Figure 1A). For more details please refer to Materials and Methods section.

We first evaluated the backbone bond lengths, bond angles, and dihedral angle distributions of the conformations generated by IDPFold. Since each residue was treated as a rigid body during the conformation generation process and parameterized using ideal relative position, we focused primarily on inter-residue features. The conformations generated by IDPFold closely matched the distributions of bond lengths and bond angles observed in MD trajectories, indicating that the model learned the chain-like arrangement between adjacent residues and accurately captured the spacing between neighboring residues (Figure 1B, Figure S1). Additionally, the distribution of Ω angles in the conformations generated by the model resembled that observed in MD trajectories, with a peak around 180°, suggesting that the local peptide bond plane conformations generated by the model are stable and align with general protein characteristics. Moreover, the Ramachandran plot composed of $\varphi - \psi$ angles showcased that the backbone dihedral angles of generated conformations predominantly fall within reasonable regions, and the density estimates across various regions closely approximate those observed in MD trajectories (Figure 1C). Specifically, with IDPFold fine-tuned on IDRome data, there was a noticeable improvement on the probability of the ppII region in the upper left corner compared to the untuned version, showing a distribution closer to MD trajectories.[13] This indicates that the fine-tuning stage enabled the model to generate more disordered structures, capturing the intrinsic biases of target proteins more precisely (Figure S2).

Next, we used IDPFold to predict conformation ensembles on four protein systems containing significant disordered regions, demonstrating the differences between the model-generated conformations and those observed in MD trajectories

(Figure 1D). The model-estimated local secondary structures and overall shape closely resembled those generated by MD, with stable conformational estimation for structured regions. As a comparison, generated conformations exhibited greater diversity in disordered regions, aligning with the highly flexible physiological characteristics of these regions.

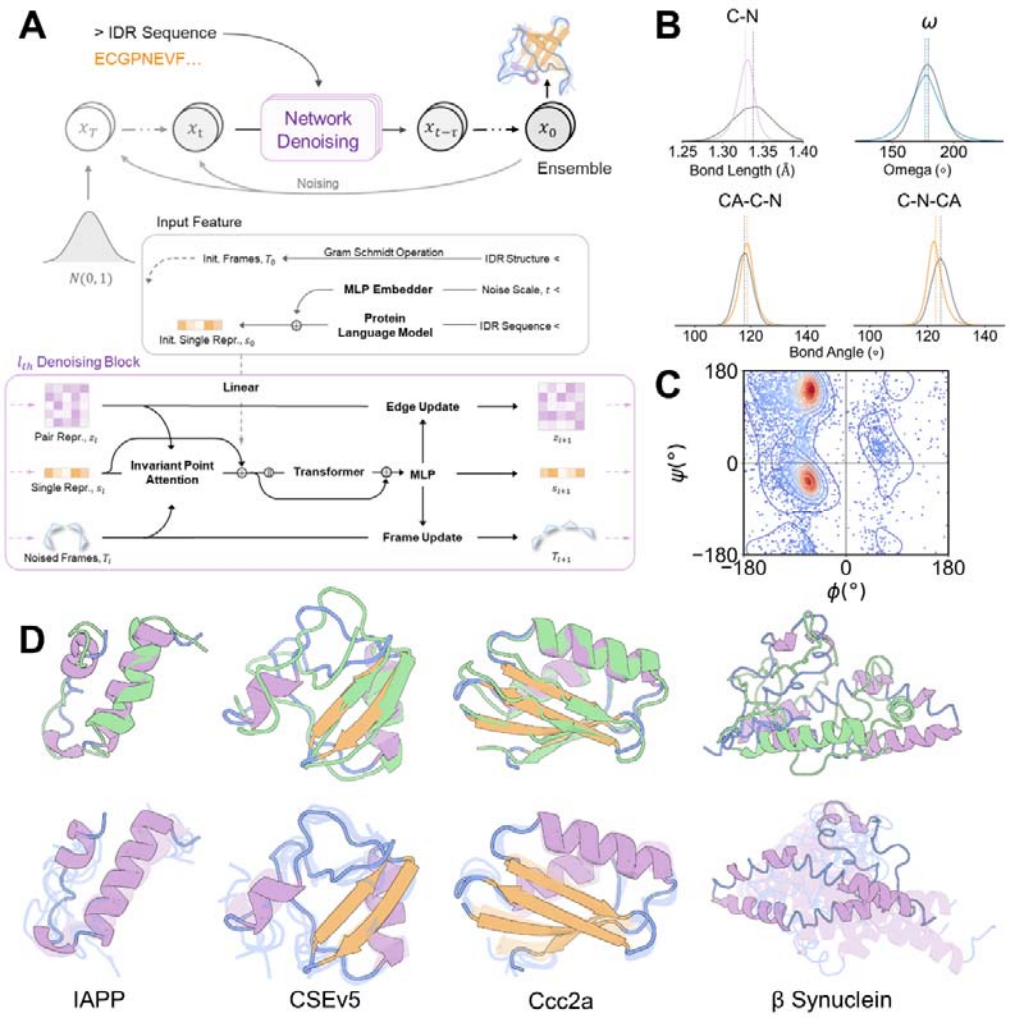

Figure 1. Overview of IDPFold. (A) Model architecture and framework for generating IDP conformations. (B) Intra-residue bond lengths, bond angles and dihedral ω of IDPFold-generated conformations compared to MD trajectories (grey). (C) Ramachandran plot, scatters are from generated samples and density lines are from MD trajectories. (D) IDPFold-generated conformations on several systems and

similar structures (green) from MD trajectories. Generated conformations are colored according to their secondary structures.

Evaluating the fidelity of IDPFold on experimental data

To assess the fidelity of the conformation ensembles predicted by IDPFold, we collected experimental physical quantities from 27 IDP systems of varying lengths (Table S1) and performed 1 μs all-atom molecular dynamics simulation for each system. All simulations were utilized with the ESFF1 force field along with the OPC3-B solvent model[18,29], which are specifically targeted for disordered proteins. Convergence of the simulation trajectories was confirmed.

We first evaluated the local features of generated ensembles using chemical shifts as the target physical quantity to measure how generated ensembles differ from experimental observations (Figure 2A). Specifically, we computed C_α secondary chemical shifts of generated conformations and the mean absolute error (MAE) between ensemble average and experimental values. C_α secondary chemical shifts can reflect the secondary structure propensity of a protein system, where positive values indicate a higher helical propensity in certain region, negative values indicate a tendency to form strands, and values close to zero indicate disordered regions. The ensembles generated by fine-tuned IDPFold are close to experimental observations, showing an average MAE of 0.48. Pairwise t-tests revealed significant improvement over MD simulations in fitting chemical shifts ($p = 0.03$), whereas pre-trained IDPFold performed worse than MD in fitting chemical shifts.

We next assessed ensemble average radius of gyration (Rg) of the generated ensembles (Figure 2B). Rg is a critical physical quantity that reflects the global characteristics of IDPs, with its magnitude positively correlating with overall protein looseness. Rg is often used to describe and distinguish IDPs from structured proteins.[29] We calculated the average error and standard deviation relative to experimental Rg values for conformation ensembles generated by IDPFold. Those generated by

IDPFold pretrained on experimental structures significantly underestimated Rg for most IDPs (average error of -22%), indicating a clear bias towards structured protein in prediction. This observation aligns with the high proportion of structured regions in experimental structures, lacking representation of natural disordered regions. However, the model significantly improved its Rg estimation after fine-tuning on MD trajectories, with an average error of only -7%. Paired t-test demonstrates a significant improvement over all-atom simulation (average error of -19%) with $p = 3.57 \times 10^{-4}$.

IDPFold not only accurately captured the ensemble average Rg of IDPs but also exhibited diversity in generated conformations comparable to MD simulations. Particularly for IDP systems with a higher structured tendency, such as ACTR and drkN SH3, the pretrained IDPFold exhibited minimal variance and mean Rg in generated structures, indicating low diversity and excessive compactness. Conversely, the fine-tuned model performed average Rg closer to experimental observations with deviations comparable to MD trajectories ($p = 0.22$ in paired t-test), suggesting both a higher diversity in generated conformations and better alignment with experimental states.

We further dived into the differences between IDPFold-predicted ensembles and experimental observations to gain a deeper understanding of the intrinsic properties of the model. Specifically, we used three specific systems as representatives, which are PaaA2, ACTR, and Sic1, respectively. From the perspective of chemical shifts, IDPFold effectively captures local secondary structure propensities in these systems (Figure 2C). In the case of PaaA2, the model accurately predicted two segments with high helical propensities in the middle and at the C-terminus, with sampled conformations showing these segments connected by a highly flexible loop region (Figure 2D). For ACTR, although the predicted ensemble closely matched experimental chemical shifts in disordered regions, it overestimated the helical propensity in the middle, resulting in predominantly stable helical segments in generated structures. This discrepancy led IDPFold to underestimate the average Rg of ACTR and the probability of disordered conformations resembling experimental

observations (Figure 2E). As for Sic1, the ensemble of generated conformations predominantly exhibited disordered states, closely matching experimental Rg values. However, given the protein's inherent β-strand propensity and the model's training with limited β-strand regions and challenges in learning long-range interactions, IDPFold generally overestimated chemical shifts in this case.

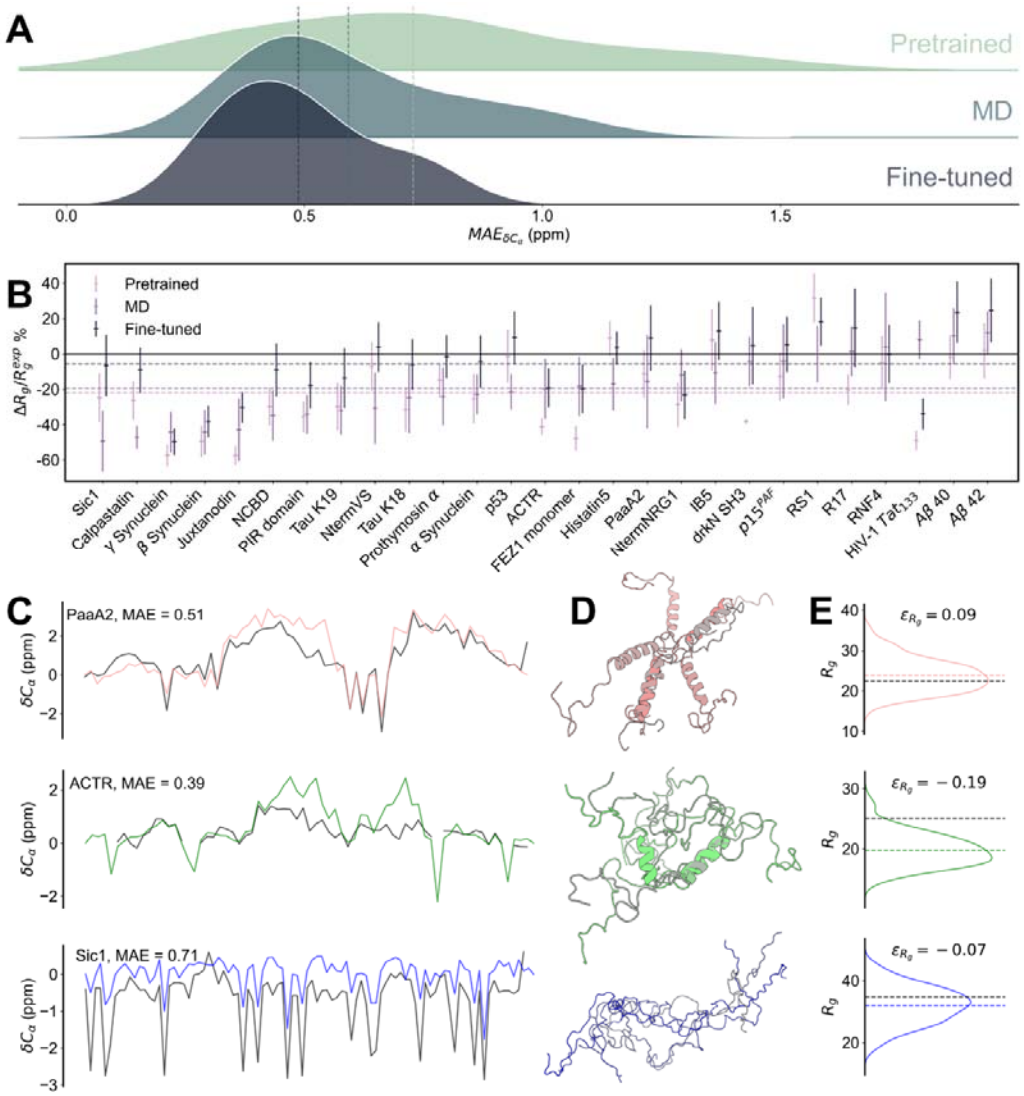

Figure 2. Fidelity of IDPFold-generated conformation ensembles evaluated on experimental data. (A) Mean absolute error on ensemble average C_α chemical shift of MD trajectories, fine-tuned and pretrained IDPFold against experimental values. (B) Ensemble average Rg error of fine-tuned IDPFold, pretrained IDPFold and all-atom

MD trajectories. (C) C_α chemical shift, (D) generated conformations and (E) Rg distributions of generated conformation ensembles on specific cases. Black lines in chemical shift and Rg are experimental data.

Comparing IDPFold to SOTA methods

We established a benchmark on 27 IDP systems and compared the performance of IDPFold with conventional MD and current state-of-the-art deep learning ensemble prediction methods. We computed the validity for each conformation ensemble as well as the fidelity composed of ε_{Rg} and $MAE_{\delta C_\alpha}$, and recorded the time required for sampling ensembles by these methods. Here, higher validity is preferred, indicating a higher proportion of reasonable distances between adjacent C_α atoms in conformations, while the value close to 0 is preferred for both ε_{Rg} and $MAE_{\delta C_\alpha}$. Together, they represent the agreement of the predicted ensemble to experimental observations. Specific definitions of these indicators can be found in the Materials and Methods section.

In terms of validity, all deep-learning based methods except EigenFold performed reasonably well. EigenFold is trained on crystal structures from the PDB dataset, in which the proportion of disordered regions is low. When transferred to IDPs, it performs poorly, barely generating reasonable protein conformations. This is also evident in its significant underestimation of Rg. For other deep learning methods, most generated conformations do not contain unreasonable C_α distances. Among them, IDPFold pretrained on experimental structures and AF-cluster exhibited the best validity, while IDPFold fine-tuned on IDP trajectories show slightly lower validity compared to the pretrained versions. This is because in MD trajectories, highly flexible IDRs often undergo significant fluctuations in atomic distances during motion, with the range of local structural changes far exceeding that of experimental structures. Therefore, after fine-tuning, the model's estimation of protein local structures tended

to fluctuate more significantly.

In terms of fidelity, the fine-tuned IDPFold performed the best in both experimental metrics. As EigenFold and idpSAM only predicted CA atomic coordinates, chemical shifts were not calculated. It is worth noting that the number of conformations produced by AF-cluster using its default clustering algorithm is uncertain, depending on the number of available MSAs and the clustering parameters. Although its generated conformations generally fit well with experimental observations, AF-cluster often fails to obtain a sufficient number of clusters in many IDP systems, producing only a few different conformations, offering little help in resolving the dynamic changes of IDPs. Moreover, the generation speed of AF-cluster was slow, taking around 1 minute to generate a single structure, making its efficiency in generating multiple conformations much lower compared to other deep learning algorithms. Meanwhile, although idpSAM-generated structures were overly compact, deviating significantly from experimental values in terms of Rg, it had a very fast inference speed, averaging only 0.63 minutes to sample 300 protein conformations. Overall, the fine-tuned IDPFold exhibited the best fit to experimental observations of IDP ensembles and its conformation generation speed was acceptable, making it highly suitable for IDP ensemble prediction tasks.

Table 1. Benchmark on IDPFold and SOTA ensemble prediction methods, bold values denote the best.

Methods	**Validity (↑)**	**Fidelity**		**Time**
		ε_{Rg}	$MAE_{\delta C_\alpha}$ (↓)	
MD $1\mu s$	-	-0.19	0.59	~200 h
AF-cluster	**0.99**	-0.12	0.57	~280 min
EigenFold	0.01	-0.62	-	21.50 min
idpSAM	0.95	-0.55	-	**0.63 min**
IDPFold, Pretrained	0.98	-0.22	0.73	20.35 min
IDPFold, Fine-tuned	0.95	$\mathbf{-0.07}$	**0.48**	

Utilizing IDPFold to sample complex thermodynamic distributions

We applied PCA to reduce the dimensionality of conformational ensembles generated by IDPFold and mapped conformations generated by AF-cluster to the same principal component space. We found that the ensemble generated by IDPFold covered the conformational space generated by AF-cluster in low-dimensional space and exhibited higher diversity (Figure 3A). As mentioned earlier, the number of conformations generated by AF-cluster for each system depends on the number of MSA clusters. In some systems with sparse MSA data, such as γ Synuclein, AF-cluster generated very few structures. For the drkN SH3 system which has rich MSA data, AF-cluster could produce an adequate number of conformations. In comparison, IDPFold sampled a similar distribution characteristic and covered a much broader range than AF-cluster. This indicates that IDPFold is able to capture the Boltzmann distribution in the protein conformational space to some extents. We also performed clustering analysis on the generated conformations of the model and visualized the top 3 cluster centroids (Figure 3B). IDPFold could sample protein conformations diverse from each other both on local secondary structures and global protein size. Although it could not provide accurate free energy estimation, we can use IDPFold to rapidly sample a large number of diverse IDP conformations as initial conformations for MD simulations to achieve broader and more accurate IDP conformation sampling.

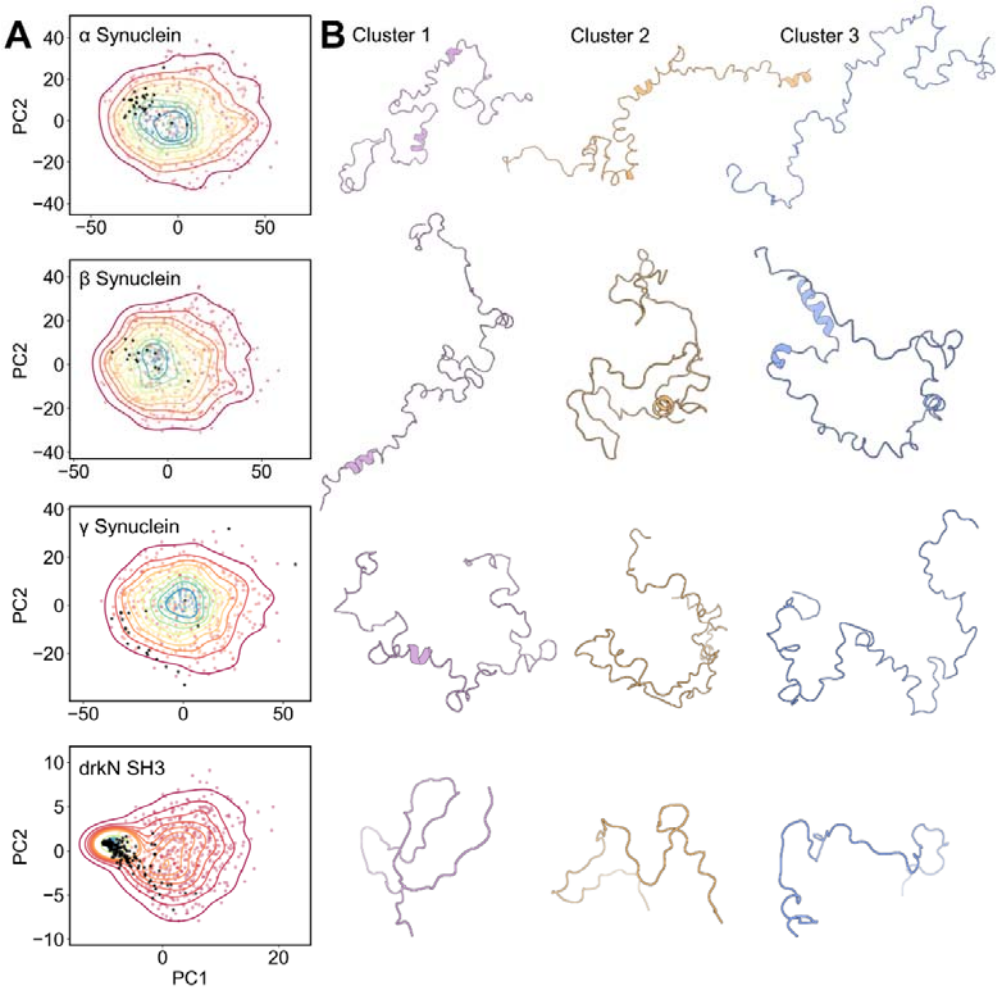

Figure 3. IDPFold provides estimation on sampled conformation distributions. (A) Principal component spaces of four generated IDP conformation ensembles, black scatters denote AF-cluster generated conformations. (B) Top 3 cluster centroids of generated conformation ensembles.

Discussion

In this study, we developed IDPFold, a deep learning-based tool for predicting ensembles of IDP conformations. IDPFold adopts a conditional diffusion model architecture to perform an end-to-end conformational generation, integrating protein language model for sequence feature extraction and DenoisingIPA module for

conformation denoising. Through a two-stage training strategy on experimental data and MD trajectories of IDPs respectively, IDPFold efficiently and accurately samples IDP conformation ensembles. Furthermore, results on the established benchmark consisting of 27 protein systems that contain IDRs and showcased the robustness of the model. IDPFold precisely capture the overall compactness and local secondary structure features of IDPs, with predicted ensembles exhibiting features with values closer to experimental observations compared to both existing MD-based and deep-learning methods. Additionally, IDPFold can sample conformational changes from structured to disordered states in proteins, demonstrating a wide sampling range and high efficiency in ensemble sampling. This capability provides important insights for studying the conformational changes and functions of IDPs.

IDPFold achieves accuracy levels comparable to, or even higher than, traditional MD simulations in estimating IDP conformation ensembles with its sampling process not restricted by energy barriers. This is not only essential for studying IDP conformations but also holds tremendous promise for dynamic proteins like allosteric proteins and enzymes, which undergo large-scale conformational changes during functional processes. Although this work focused on training the model for IDPs and direct prediction performance on allosteric protein ensembles may not be optimal, the methodology would likely to be transferred on diverse categories of proteins, enabling more accurate estimations of conformational transitions in these functionally important proteins.

Although IDPFold demonstrates robustness and precision in predicting conformation ensembles, its inference time is relatively long compared to some of the existed deep learning methods with an average time of about 20 minutes to sample an entire system. While this speed is much faster than conventional MD simulations, there is room for further improvement on model efficiency. To achieve higher sampling accuracy and more precisely protein backbone features, we employed a more complex network architecture and longer diffusion steps which might well sacrifice sampling efficiency. To enhance the inference efficiency of the model,

optimizations can be explored such as reducing the number of diffusion steps in the diffusion model and optimizing transformer components by replacing them with architectures that are more efficient in terms of time and space. Additionally, in this study, we only trained and tested IDPFold on single-chain proteins and did not evaluate it on protein complexes or multi-chain proteins. In future work, we aim to extend the model to these areas to efficiently sample and interpret how molecules dynamically interact with each other.

Conformational ensemble prediction represents a significant frontier in protein structure research following static structure prediction. Accurate characterization of protein conformational ensembles can help us understand the dynamics of proteins and crucial conformational changes they undergo when binding to substrates or undergoing biochemical reactions under physiological conditions. Although deep learning-based tools are not yet widely adopted and most dynamic analyses still relying on MD simulations, our work suggest that properties predicted using deep learning methods may offer higher accuracy than MD simulations, potentially serving as complementary or even alternative approaches to MD simulations. However, current deep learning-based conformational ensemble prediction methods still face several challenges, such as inaccurate estimation of conformational free energies, less robustness compared to traditional methods, and the inability to capture temporal autocorrelation between conformations (i.e., inability to capture dynamic features).

Several studies have indicated that the single-point energy estimation of molecular force fields can be used to train diffusion models, allowing models to learn more comprehensively about conformational space in cases where structural data are scarce[30,31]. This training approach, which utilizes force field energies rather than structural data, enables more accurate capture of conformational space distribution and greater robustness. However, this training strategy also makes the model sensitive to empirical parameters of the force field, demanding high precision in molecular force field accuracy. Additionally, there are works based on flow models or score-matching models aiming to fit the dynamic characteristics of MD trajectories, which

may be a solution to the current inability of generative models to capture temporal autocorrelation between conformations[32].

Apart from the challenges of sparse training data and limited model representation capabilities, there is another significant issue in conformational ensemble prediction tasks: the scarcity and lack of uniformity in evaluation metrics. While we collected a set of experimental observations as a gold standard for this study, a vast number of protein systems lack experimental annotations, making it challenging to assess the generative performance of computational models on these systems. Therefore, we believe that constructing confidence metrics for conformational ensemble generation akin to pLDDT for protein structure prediction is also promising for evaluating and improving protein conformational ensemble models.

Materials and Methods

Datasets

The data for training IDPFold consists of three components: high-quality crystal structures collected from the PDB dataset, conformations ensembles resolved by NMR and molecular dynamics (MD) simulation dataset of IDPs obtained through extensive back-mapping and energy minimization on coarse-grained trajectories.

For crystal structures, we referenced trRosetta and gathered a total of 15,051 X-ray resolved structures with resolutions ≤ 2.5Å and sequence redundancy $\leq 30\%$[33]. These structures provide fundamental protein characteristics for the model, such as bond lengths, bond angle distributions and chirality of C_α. However, crystal structures typically depict stable conformations of proteins, i.e., structured conformations. Training the model solely on crystal structures would significantly underestimate the disorder tendency of IDPs and result in low diversity in conformation generation for individual protein sequences. Therefore, we additionally collected 12,339 NMR resolved protein conformation ensembles from the PDB. We filtered these NMR

entries based on 30% sequence similarity both internally and against crystal structures, resulting in 539 systems comprising a total of 10,454 structures. By blending crystal structures with NMR ensembles, we obtained a combined total of 25,495 experimental structures for the initial phase of IDPFold training. These structures were well-defined and exhibited a higher structured tendency, ensuring that the model captures the local physical characteristics of proteins effectively.

For the IDP trajectory data, we obtained large-scale coarse-grained simulation data from IDRome that recorded most IDRs in human proteome[21]. These coarse-grained simulations accurately capture global features of IDPs (and IDRs), such as average radius of gyration (Rg) of conformation ensembles, but they only retain C_α atoms, whose resolutions fail to meet the requirement for model training. Therefore, we selected all systems with lengths exceeding 256 residues, totaling 3,880 systems. Choosing larger systems is primarily because these long disordered segments encompass most sequence and structural characteristics found in smaller systems, and training the model on larger systems can help enhance its generalization ability. For these systems, we first used pdbfixer to restore structures from coarse-grained to all-atom and then performed 100 steps of energy minimization with ff14SB[34] force field for all protein structures. Evaluation showed that this optimization process effectively eliminated unreasonable conformations in the back-mapped structures (Figure S1). A total of 77,600 optimized all-atom conformations was used for the second phase of IDPFold training, these conformations globally exhibit more disordered and can aid the model in learning distinctive conformational features of natural IDPs.

Furthermore, data used to assess the performance of IDPFold generation included 27 IDP systems that are fully described by experiments and are not present in the training set. We conducted $1\mu s$ all-atom MD simulations for these systems with the IDP-specific force field ESFF1 and solvent model OPC3-B[18,29]. These all-atom simulation trajectories described the dynamics and thermodynamic characteristics of IDPs, which are suitable for evaluating the quality of IDPFold-generated conformational ensembles.

Formulation of IDPFold

Diffusion Modeling on Protein Structure

To enable IDPFold to capture Boltzmann distribution of protein conformations at equilibrium state, we employ Score-Based Generative Modeling (SGM) to learn the probability distribution from protein structure data. SGM can be represented by a diffusion process $x_t \in \mathbb{R}^n$ defined by stochastic differential equation (SDE)[35]. The forward diffusion process is characterized by the following equation:

$$dx = f(x,t)dt + g(t)dw \#(1)$$

where $t \in [0,T]$ is a continuous index and $w \in \mathbb{R}^n$ is the standard Wiener process (a.k.a., Brownian motion). $f(x,t) \in \mathbb{R}^n$ is a vector-valued function called the drift coefficient, and $g(t) \in \mathbb{R}$ is a scalar function called the diffusion coefficient. Then, the corresponding backward diffusion process, or denoising process can also be defined by SDE[36]:

$$dx = [f(x,t) - g^2(t)\nabla_x logp_t(x)]dt + g(t)d\overline{w} \#(2)$$

where $d\overline{w}$ is a standard Wiener process as continuous time t flows backwards from T to 0, and dt is an infinitesimal negative time step. $x_{t=0}$ here represents ground truth data, or the protein conformations, while $x_{t=T}$ is sampled from Gaussian distribution. Therefore, by solving the backward SDE process, we can sample diverse protein conformations that obey Boltzmann distribution from a Gaussian distribution. In equation 2, each term except the score of x, $\nabla_x logp_t(x)$, is solvable. Therefore, we only needed a score-matching network to fit $\nabla_x logp_t(x)$ at each time step to achieve the purpose of generating conformations.

Based on the above-mentioned standard formulation of SGM, we further required that the entire conformation generation process should be SE(3)-equivariant, i.e., diffusion process and network transformation were not sensitive to global rotation and translation of protein structures. SE(3)-equivariance can be described by the following equation:

$$\mathcal{F} \circ \rho(x) = \rho \circ \mathcal{F}(x) \#(3)$$

where $\mathcal{F}$ denotes data transformations like network prediction and diffusion process, ρ represents global rotation and translation. Typically, a protein conformation x is characterized by Cartesian coordinates $c_i \in \mathbb{R}^3, 1 \leq i \leq N$, where N denotes number of atoms. However, data transformations on Cartesian coordinates is computationally intensive and does not easily satisfy SE(3)-equivariance. Therefore, we adopted backbone frame parametrization, representing the protein conformation x as $T_j \coloneqq [R_j, v_j], 1 \leq j \leq n$, where n denotes the number of residues. Each backbone frame includes a 3×3 rotation matrix $R_j \in SO(3)$ and a translation vector $v_j \in \mathbb{R}^3$. A frame T_j can represent Euclidean transformation for each atom in residue j from local coordinates c_{local} to global coordinates c_{global} as $c_{global} = T_i \circ c_{local} \coloneqq R_i\, c_{local} + v_i$. Following the approach described in FrameDiff[31], we independently handle the rotation matrix R_j on the $SO(3)$ manifold and the translation vector v_j in $\mathbb{R}^3$ during the diffusion process, formulated as follows:

$$dT_t = \left[0, -\frac{1}{2}\beta(t)Pv_t\right]dt + \left[\sqrt{\frac{d}{dt}\sigma^2(t)}dw^{(SO(3))}, \sqrt{\beta(t)}Pdw^{(\mathbb{R}^3)}\right] \#(4)$$

Where $\beta(t)$ and $\sigma(t)$ control the scale of noise during the diffusion process, $w^{\mathcal{M}}$ denotes Brownian motion defined on manifold $\mathcal{M}$, and $P: \mathbb{R}^{3n} \to \mathbb{R}^{3n}$ is used for removing the center of mass. During the forward diffusion or noising process, the addition of noise on rotation matrices is determined by the noise kernel $p_{t|0}(R_t|R_0)$, which is obtained from an isotropic Gaussian distribution on the $SO(3)$ manifold. This distribution is formulated as:

$$\mathcal{IG}_{SO(3)}(R_t; R_0, \sigma^2) = f(\omega_{t|0}) \coloneqq \frac{1-\cos(\omega_{t|0})}{\pi}\sum_{l=0}^{\infty}(2l+1)e^{-l(l+1)\sigma^2\frac{\sin((l+0.5)\omega_{t|0})}{\sin(0.5\omega_{t|0})}} \#(5)$$

Here, $\omega_{t|0} = Axis_angle(R_0^T R_t)$ is the axis-angle transformed representation of the composed rotation matrix $R_0^T R_t$. As for the translation vector, its noise addition process follows an Ornstein-Uhlenbeck process, also known as VP-SDE. The noise kernel for the translation vector is relatively straightforward, converging ultimately to

$\mathcal{N}(0, I)$ as shown in the following equation.

$$p_{t|0}(v_t|v_0) = \mathcal{N}\left(v_t; v_0 e^{-\frac{1}{2}\int_0^t \beta(s)ds}, I - Ie^{\int_0^t \beta(s)ds}\right) \#(6)$$

Network Design and Training

To achieve the goal of predicting conformation ensembles from sequence, we devised a sequence-conditioned score-matching model $s_\theta(x_t, x, seq)$ for denoising processes. As the diffusion process is designed to be strictly SE(3)-equivariant, network transformations should evidently preserve this property. Therefore, we adopted a variant of structure module from AlphaFold2 (Figure 1A) for updating backbone frames[22]. Here, we employed the Point Invariant Attention (IPA) mechanism to capture interactions and relationships between nearby residues, followed by a Transformer to learn global features and long-range interactions. This architecture has been demonstrated in previous research to promote training and generation of high-quality protein conformations. The aforementioned network design requires three inputs foreach layer: a one-dimensional vector representation s_l, pairwise feature representation z_l, and the set of rotation and translation updates T_l. We utilized ESM2-650M to extract protein sequence features, concatenated with residue position encoding and time encoding represented by trigonometric functions, to form the initial one-dimensional vector representation s_0. Pairwise feature representation z_0 is derived from s_0 based on relative positional encoding. After each layer of IPA-Transformer transformation, we updated the one-dimensional vector representation through a fully connected network and subsequently updated the pairwise features *via* cross product.

The objective of score-matching networks differs from conventional neural network training goals. It does not aim to fit protein conformations directly but rather the scores of data perturbed to a certain degree, i.e., fitting the distribution of perturbed data. To measure how well the predicted scores fit actual distribution, we computed the DSM loss as follows:

$$\mathcal{L}_{dsm} = \mathbb{E}_{t\in[0,T]}\left\{\lambda(t)\mathbb{E}_{T_0}\mathbb{E}_{T_t|T_0}\left[\left|\left|s_\theta(T_t,t) - \nabla_{T_t} log p_{t|0}(T_t|T_0)\right|\right|^2\right]\right\} \#(7)$$

To ensure that the DSM loss at all time steps t results in a perfect fit score of 1, ensuring equal contribution of each time step to the loss function, we set the weights as follows:

$$\lambda(t) = \frac{1}{\mathbb{E}\left[\left\|\nabla_{T_t} log p_{t|0}(T_t|T_0)\right\|\right]} \#(8)$$

Additionally, to ensure the model learns detailed features of protein structures, apart from the DSM loss on rotation and translation matrices, we also incorporate mean square error (MSE) supervision for the positions of backbone atoms and differences on the distance matrix for samples with fewer forward diffusion steps ($t < \frac{T}{4}$). Therefore, the complete network training loss function can be represented as:

$$\mathcal{L} = \mathcal{L}_{dsm} + \omega_1\mathcal{L}_{bb} + \omega_2\mathcal{L}_{dist} \#(9)$$

where $\omega_1 = \omega_2 = 0.25$ control the weight of conformation quality loss.

During the training process, with a maximum time step $T = 1.0$, we optimize the DenoisingIPA using an Adam optimizer with the learning rate of 10^{-4}[37]. For the translation vector part of the network training, we employ a linear noise strategy within the VP-SDE framework, while for the rotation matrix part, we use a logarithmic noise strategy within the VE-SDE framework, as shown in the following equation[35]:

$$\beta(t) = \beta_{min} + \frac{t}{T}(\beta_{max} - \beta_{min}),$$
$$\beta_{min} = 0.1, \beta_{max} = 20 \#(10)$$

$$\sigma(t) = \log(te^{\sigma_{max}} + (T-t)e^{\sigma_{min}}),$$
$$\sigma_{min} = 0.1, \sigma_{max} = 1.5\#(11)$$

Implementation Details

The training of IDPFold consists of two stages: pre-training on experimental structures and fine-tuning on MD trajectories. In Table 1, we present the model hyperparameters utilized during the 2 training stages:

Table 1. Hyperparameters and training details of IDPFold

Hyperparameters	Training Stage	
	Training on Experimental Data	Training on MD Trajectories
Single Repr. Channel	256	
Pair Repr. Channel	128	
Hidden Channel	256	
IPA Layers	4	
Transformer Layers	2	
Transformer Heads	8	
Model Size	17.8 M parameters	
Learning Rate	10^{-4}	10^{-5}
Batch Size	8	32
Iterations	1.12 M	0.44 M
Time	~9 GPU days	~15 GPU days

Evaluation Metrics and Analysis Tools

We evaluated the quality of IDPFold-predicted IDP conformational ensembles from two perspectives: local features of the generated structures and global features of the conformational ensembles. For local features, biotite is used to compute inter-residue bond lengths and bond angles of all generated conformations, specifically, the $C-N$ bond lengths, $C_\alpha - C - N$ and $C - N - C_\alpha$ bond angles[38]. We also analyzed the distribution of backbone dihedral angles to assess the model's prediction accuracy regarding C_α chirality and secondary structure. Additionally, we used mdtraj to calculate scalar coupling between HN and H_α, employed SPARTA+ to compute C_α chemical shifts, and then performed ensemble averaging[39,40]. These analyses helped determine whether the local environment of the protein backbone aligns with experimental observations.

For global features, we calculated Rg of the generated conformations and

compared the ensemble average with experiment results. Additionally, we derived C_α distance matrices to facilitate PCA dimensionality reduction. Through clustering the generated conformations and projecting them onto principal component space, we aim to discern the diversity of generated conformations and assess the model's learning efficacy in capturing the Boltzmann distribution information from MD trajectories in the training dataset.

Due to the fact that most current work on generating protein conformational ensembles from sequences adopts a coarse-grained representation, where only C_α coordinates are generated, we evaluated Validity and Fidelity at this coarse-grained level and compared the performance of IDPFold with previous methods.

Validity assesses whether the generated conformations contain unreasonable C_α distances. In a protein structure, due to van der Waals interactions between atoms, C_α distances should not be too close. Additionally, because neighboring C_α in the protein backbone are connected by $C - N$ with specific bond lengths, C_α distances should not be excessively far apart. Therefore, following the approach of Str2Str, we defined a reasonable range for C_α distances as follows:

$$\delta_{vdw} < d_{C_\alpha^i, C_\alpha^{i+1}} < \delta_{bond} \#(12)$$

where $\delta_{vdw} = 2 \times 1.7 - 0.4$ is defined as the sum of two C_α van der Waals radius minus an acceptable overlap distance of 0.4[41]. δ_{bond} is taken as the maximum C_α distance observed in MD trajectory of each test system. Since a reasonable range is defined, we could define validity as proportion of valid conformations. A higher validity indicates a lower probability of mis-estimated C_α distances, thus demonstrating better performance.

Fidelity refers to how well the model-generated ensembles match experimental observations. We selected Rg and C_α secondary chemical shifts as target physical quantities to measure the fidelity of models, calculating the errors between model-generated conformation ensembles and experimental observations using the following formula:

$$\varepsilon_{Rg} = \frac{\langle Rg \rangle^{pred} - \langle Rg \rangle^{exp}}{\langle Rg \rangle^{exp}} \#(13)$$

$$MAE_{\delta C_\alpha} = \mathbb{E}(|\langle \delta C_\alpha \rangle^{pred} - \langle \delta C_\alpha \rangle^{exp}|) \#(14)$$

where $\langle Rg \rangle$ denotes the ensemble average Rg, $\langle Rg \rangle^{pred}$ is the generated ensemble average Rg while $\langle Rg \rangle^{exp}$ is experimental observation. Similarly, $\langle \delta C_\alpha \rangle^{pred}$ and $\langle \delta C_\alpha \rangle^{exp}$ denotes ensemble average chemical shift and experimental observation respectively. Using the above definitions, we calculated ε_{Rg} and $MAE_{\delta C_\alpha}$, aiming for values close to zero for both. ε_{Rg} reflects the reasonableness of the model's generated conformations in terms of their compactness, while $MAE_{\delta C_\alpha}$ indicates how well the generated conformations conform to experimental observations on local secondary structures.

For evaluating the conformational ensembles generated by the models, all methods except AF-cluster generated 300 conformations for subsequent evaluation[42,43]. This number of conformations has been previously shown to adequately reflect the structural diversity of intrinsically disordered proteins. The number of conformations generated by AF-cluster is influenced by the number of clusters in the MSA, and we used its default MSA and clustering settings. All conformation generation was performed on NVIDIA A100 GPUs.

Acknowledgements

This work was supported by the Center for HPC at Shanghai Jiao Tong University, and the National Key Research and Development Program of China (2023YFF1205102 and 2020YFA0907700), the Fundamental Research Funds for the Central Universities (YG2023LC03), and the National Natural Science Foundation of China (21977068, 32171242 and T2222012).

Data Availability

The Supplementary Information is still in progress, the dataset and code of IDPFold will be available at GitHub.

References

1. Dunker, A. K. *et al.* The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics* **9**, S1 (2008).
2. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradović, Z. Intrinsic Disorder and Protein Function. *Biochemistry* **41**, 6573–6582 (2002).
3. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Intrinsically Disordered Proteins in Human Diseases: Introducing the D ^{2} Concept. *Annu. Rev. Biophys.* **37**, 215–246 (2008).
4. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradović, Z. & Dunker, A. K. Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins. *J. Mol. Biol.* **323**, 573–584 (2002).
5. Zhao, R. *et al.* Analysis of p53-regulated gene expression patterns using oligonucleotide arrays. *Genes Dev.* **14**, 981–993 (2000).
6. Singleton, A. B. *et al.* α-Synuclein Locus Triplication Causes Parkinson's Disease. *Science* **302**, 841–841 (2003).
7. Casu, F., Duggan, B. M. & Hennig, M. The Arginine-Rich RNA-Binding Motif of HIV-1 Rev Is Intrinsically Disordered and Folds upon RRE Binding. *Biophys. J.* **105**, 1004–1017 (2013).
8. Fernández, A. & Scheraga, H. A. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 113–118 (2003).
9. Chruszcz, M., Borek, D., Domagalski, M., Otwinowski, Z. & Minor, W. X-ray diffraction experiment – the last experiment in the structure elucidation process. *Adv. Protein Chem. Struct. Biol.* **77**, 23–40 (2009).
10. Dandey, V. P. *et al.* Time-resolved cryoEM using Spotiton. *Nat. Methods* **17**, 897–900 (2020).
11. McDermott, A. Structure and Dynamics of Membrane Proteins by Magic Angle Spinning Solid-State NMR. *Annu. Rev. Biophys.* **38**, 385–403 (2009).
12. Jensen, M. R., Zweckstetter, M., Huang, J. & Blackledge, M. Exploring Free-Energy Landscapes of Intrinsically Disordered Proteins at Atomic Resolution Using NMR Spectroscopy. *Chem. Rev.* **114**, 6632–6660 (2014).
13. Huang, J. *et al.* CHARMM36m: An Improved Force Field for Folded and

Intrinsically Disordered Proteins. *Nat. Methods* **14**, 71–73 (2017).
14. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4758–E4766 (2018).
15. Li, Z., Mu, J., Chen, J. & Chen, H.-F. Base-specific RNA force field improving the dynamics conformation of nucleotide. *Int. J. Biol. Macromol.* **222**, 680–690 (2022).
16. Zhang, Y., Liu, H., Yang, S., Luo, R. & Chen, H.-F. A Well-balanced Force Field ff03CMAP for Folded and Disordered Proteins. *J. Chem. Theory Comput.* **15**, 6769–6780 (2019).
17. Song, D. *et al.* ff14IDPs Force Field Improving the Conformation Sampling of Intrinsically Disordered Proteins. *Chem. Biol. Drug Des.* **89**, 5–15 (2017).
18. Song, D., Liu, H., Luo, R. & Chen, H.-F. Environment-Specific Force Field for Intrinsically Disordered and Ordered Proteins. *J. Chem. Inf. Model.* **60**, 2257–2267 (2020).
19. Zhong, B., Song, G. & Chen, H.-F. Balanced Force Field ff03CMAP Improving the Dynamics Conformation Sampling of Phosphorylation Site. *Int. J. Mol. Sci.* **23**, 11285 (2022).
20. Mu, J., Pan, Z. & Chen, H.-F. Balanced Solvent Model for Intrinsically Disordered and Ordered Proteins. *J. Chem. Inf. Model.* **61**, 5141–5151 (2021).
21. Tesei, G. *et al.* Conformational ensembles of the human intrinsically disordered proteome. *Nature* **626**, 897–904 (2024).
22. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
23. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
24. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
25. Ingraham, J. B. *et al.* Illuminating protein space with a programmable generative model. *Nature* **623**, 1070–1078 (2023).
26. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
27. Lazar, T. *et al.* PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res.* **49**, D404–D411 (2020).
28. Riley, A. C., Ashlock, D. A. & Graether, S. P. The difficulty of aligning intrinsically disordered protein sequences as assessed by conservation and phylogeny. *PLOS ONE* **18**, e0288388 (2023).
29. Pan, Z., Mu, J. & Chen, H.-F. Balanced Three-Point Water Model OPC3-B for Intrinsically Disordered and Ordered Proteins. *J. Chem. Theory Comput.* **19**, 4837–4850 (2023).
30. Zheng, S. *et al.* Towards Predicting Equilibrium Distributions for Molecular Systems with Deep Learning. Preprint at https://doi.org/10.48550/arXiv.2306.05445 (2023).
31. Yim, J. *et al.* SE(3) diffusion model with application to protein backbone

generation. Preprint at http://arxiv.org/abs/2302.02277 (2023).
32. Noé, F., Olsson, S., Köhler, J. & Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **365**, eaaw1147 (2019).
33. Yang, J. *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci.* **117**, 1496–1503 (2020).
34. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
35. Song, Y. *et al.* Score-Based Generative Modeling through Stochastic Differential Equations. Preprint at http://arxiv.org/abs/2011.13456 (2021).
36. Anderson, B. D. O. Reverse-time diffusion equation models. *Stoch. Process. Their Appl.* **12**, 313–326 (1982).
37. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. Preprint at http://arxiv.org/abs/1412.6980 (2017).
38. Kunzmann, P. & Hamacher, K. Biotite: a unifying open source computational biology framework in Python. *BMC Bioinformatics* **19**, 346 (2018).
39. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109**, 1528–1532 (2015).
40. Shen, Y. & Bax, A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR* **48**, 13–22 (2010).
41. Ramachandran, S., Kota, P., Ding, F. & Dokholyan, N. V. Automated Minimization of Steric Clashes in Protein Structures. *Proteins* **79**, 261–270 (2011).
42. Daughdrill, G. W. *et al.* Understanding the Structural Ensembles of a Highly Extended Disordered Protein. *Mol. Biosyst.* **8**, 308–319 (2012).
43. Schwalbe, M. *et al.* Predictive Atomic Resolution Descriptions of Intrinsically Disordered hTau40 and α-Synuclein in Solution from NMR and Small Angle Scattering. *Structure* **22**, 238–249 (2014).